EXPERIMENTS ON LANGUAGE ERRORS IN AVIATION MAINTENANCE

Colin G. Drury and Jiao Ma
University at Buffalo, Department of Industrial Engineering
438 Bell Hall, Buffalo, NY 14260
drury@buffalo.edu

The Federal Aviation Administration has raised many issues concerning the outsourcing of maintenance to foreign repair stations and recommends establishing a method for determining whether language barriers result in maintenance deficiencies. This work addresses concerns that non-native English speakers may be prone to an increased error rate that could potentially affect airworthiness. This paper presents Year 2 of the project. We used the seven scenarios of language error developed in Year 1 as the basis for our data collection effort to quantify the frequency of error. An intervention experiment has been designed and tested on two groups of participants: English-speaking maintenance personnel and Chinese speaking engineering graduate students. Neither is the final target group, but the methodology needed to be verified before on-site data collection. The analysis of the data from these two experiments is presented here.

INTRODUCTION

In 2001, the Federal Aviation Administration raised many issues concerning the outsourcing of maintenance to foreign repair stations in considering changes to domestic and foreign Federal Air Regulations, recommending that:

> *"The FAA should establish a method for determining whether language barriers result in maintenance deficiencies."*

This project is a direct response to these concerns that non-native English speakers, in repair stations in the USA and abroad, may be prone to an increased error rate that could potentially affect airworthiness. The documentation for repair provided by an English speaking airline is always in English, and this documentation must be used to govern all maintenance tasks, despite a potentially large proportion of mechanics who do not use English as a native language. This paper follows our 2003 HFES paper (Drury and Ma, 2003) and describes two experiments testing a methodology for quantifying the effectiveness of possible countermeasures to language errors.

As noted in our 2003 paper, this project developed seven scenarios of language error based on visits to sites in the USA and the UK; it also provided a model for these unique communication errors based on the communications literature and an analysis of several databases. These included the NASA/ASRS error database and responses to a questionnaire on language skills provided by a major manufacturer. Our analyses showed that language skill varied (as expected) by world region, and that not all sites with lower language skills translated documents into the native language. Many references to communication theories and studies of outsourcing were given in Drury and Ma (2003) and will not be repeated here.

The seven scenarios found were:

Scenario 1: "The Mechanic (AMT) or Inspector was not able to communicate verbally to the level required for adequate performance."
Scenario 2: "The Mechanic (AMT) or Inspector and the person to whom they were speaking did not realize that the other had limited English ability."
Scenario 3: "Native English speakers with different regional accents did not understand each others' communications."
Scenario 4: "The Mechanic (AMT) or Inspector did not understand a safety announcement over the Public Address (PA) system."
Scenario 5: "The Mechanic (AMT) or Inspector did not fully understand a safety placard."
Scenario 6: "The Mechanic (AMT) or Inspector did not fully understand documentation in English, for example a Work Card or a Manual."
Scenario 7: "The Mechanic (AMT) or Inspector did not fully understand a document translated from another language into their native language."

In our continuing work, we will be visiting sites worldwide to measure the frequency of these scenarios, but the current paper concentrates on the second aspect of the work, that of evaluating countermeasures.

Our analysis of worldwide survey data from a major manufacturer reported earlier found that two strategies used to reduce the potential for language errors were (a) translation into the native language, and (b) conducting face-to-face meetings in the native language. However, only about 17% of airlines in the region that most often used translation (Asia) actually translated maintenance documents into the native languages. Even among the group of 8 airlines who reported the lowest English speaking ability, only 2 modified the English documents in any way. Other strategies of intervention found in our site visits included having a bilingual English/native language speaker assist the mechanic with the English documentation, and/or providing a glossary of key words between the native language and English. Finally, our own earlier research into the artificial maintenance language called AECMA Simplified English (e.g., Chervak, Drury and Ouellette, 1996) had shown this to be an effective error reduction technique, particularly for non-native English speakers and for complex work documents.

Thus, we will compare four potential language error reduction interventions:

- The translation of a document into AECMA Simplified English
- The provision of a Glossary
- The provision of a bilingual coach
- The translation of a document and all related materials into a native language

Some of these methods may be combined, for example the provision of both a Glossary and a bilingual coach, or the addition of AECMA Simplified English to all conditions except for translation into the native language. Finally, for comparison, a baseline condition, no intervention, will be required. This paper describes the first two experiments conducted within this framework.

METHODOLOGY

Measures

To test for how potential documentation errors can be reduced, we measured the effectiveness of document comprehension. In the study, a single task card was given to participants with a 10-item questionnaire to test comprehension. The methodology was validated in our previous research (e.g., Chervak, et al., 1996; Drury, Wenner and Kritkausky, 1999). The comprehension score was measured by the number of correct responses, with time taken to complete the questionnaire as an additional measure.

Task Cards

We selected two task cards, one "easy" and one "difficult," from four task cards used in the previous research, because it had already been found that task difficulty affected the effectiveness of one strategy, Simplified English. As was expected, the use of Simplified English had a larger effect on more complex task cards (Chervak and Drury, 2003). The complexity of these task cards was evaluated by Boeing computational linguists and University of Washington technical communications researchers considering word count, words per sentence, percentage passive voice and the Flesch-Kincaid reading score. The cards differed on all measures.

Both of the task cards were then prepared in the European Association of Aerospace Industries (AECMA) Simplified English versions, which were also critiqued by experts from Boeing, the University of Washington, and the American Institute of Aeronautics and Astronautics (AIAA) Simplified English Committee. We also used a short test of English ability, the Accuracy Level Test (Carver, 1987), to act as a potential covariate in our analysis.

Design

As shown in **Table 1**, our study is a three-factor factorial design with the participants nested under the three factors of:

a. Task Card Complexity (Easy vs. Difficult)
b. Document Language (Simplified English vs. Non-simplified English)
c. Interventions (None, Glossary, Full Translation, Bilingual Coach, Glossary Plus Bilingual Coach)

Intervention	**Easy Task Card**		**Difficult Task Card**	
	Simplified English	**Non-Simplified English**	**Simplified English**	**Non-Simplified English**
	#2 **#1**	**#2** **#1**	**#2** **#1**	**#2** **#1**
1. Control	**2** **4**	**2** **3**	**2** **4**	**2** **4**
2. Glossary	**2**	**2**	**2**	**2**
3. Tutoring	**2**	**2**	**2**	**2**
4.Glossary & Tutoring	**2**	**2**	**2**	**2**
5.Chinese Translation	**2**	**2**	**2**	**2**

NOTE: #1 represents the number of participants in Pilot Test 1, and #2 represents the number of participants in Pilot Test 2.

Table 1. Participant Numbers by Experimental Conditions for Pilot Tests 1 and 2

Choice of Participants and Sites

The main task will take place at various foreign Maintenance/Repair organizations (MROs), but the two studies reported here were performed in the USA as baseline and pilot tests.

There are several reasons to collect data from MROs located in Asia, especially China. First, in our analysis of the manufacturer's survey data, we found that about 30% of users in Asia had a very limited English speaking ability, another 40% were able to conduct simple conversations; about 40% of the users were able to work effectively with only written maintenance/inspection related documents, and another 15% had very little English reading ability.

Compared with North America and Europe, Asia has a much smaller base of English-using mechanics. Second, the Asia-Pacific region is poised to be one of strongest growth engines for the foreseeable future for the maintenance, repair and overhaul industry (*Overhaul & Maintenance*, 2002). U.S. and European airlines continue to ship wide-body aircraft to East Asia to take advantage of low labor costs. Almost half of the top 10 Asian MROs are located in China. According to *Aviation Week & Space Technology*, "the Civil Aviation Administration of China (CAAC) is confident that despite the downturn in the global airline industry, more maintenance, repair and overhaul (MRO) joint venture companies will be set up with Chinese airlines within the next two years" (Dennis, 2002).

Participants were tested singly or in small groups. After obtaining Informed Consent and completing demographic questions, participants were given one of the four task cards and its associated comprehension questions. They were timed, but instructions emphasized accuracy. After the completion of the comprehension task, participants were given the Accuracy Level Test for the required 10 minutes. This test used a total of 100 words with a forced synonym choice among three alternatives, and produced on the scale of reading grade level. It has been validated against more detailed measures of reading level (Chervak, Drury, Ouellette, 1996).

The Preparation of the Data Collection Packet in Chinese

The translation process took place in two steps. A native Chinese research assistant (9 years as an engineering major), who is very familiar with the task cards, took a lead in translating the packet. A large number of technical and language references were consulted. The principal investigator and other domain experts (e.g., native Chinese mechanical engineers in the Department of Aerospace and Mechanical Engineering at the University at Buffalo, SUNY) were consulted on the technical details (e.g., lockwire). Then both translated the task cards, and original packets of data collection material were submitted to a retired professor from the Department of Avionics, Civil Aviation University of China (CAUC) for a review.

We developed an English/Chinese glossary for each task card. We had two native English speaking engineering graduate students and two native Chinese speaking engineering graduate students read through all the task cards and circle all the words/phrases/sentences they did not comprehend, or even those about which they were slightly unsure. We developed this glossary to be as comprehensive as possible, including nouns, verbs, adjectives, abbreviations, etc.

With all of this material prepared, we performed two experiments before visiting the Chinese-speaking sites.

The Results of Pilot Test 1: Native English speaking Maintenance Personnel

This test used 15 participants from three sites in the UK and the USA as a representative sample of English-speaking maintenance personnel who were unlikely to have any language errors. They were tested on the same visits where focus group data was collected, as reported in Drury and Ma (2003). All were tested under the four combinations of Simplified English/Not and Easy/Difficult Task Card to give a 2 x 2 between subjects design. There were no other interventions with these native English speakers.

First, there was a high negative correlation between accuracy and time for the comprehension test (r = 0.692, p = 0.004), and moderate correlations of both with Reading Level at p = 0.06. Thus, another variable was created through dividing Accuracy by Time to give a combined overall Performance score. Reading Level was tested as a covariate, but was not significant in any of three GLM ANOVAs of Accuracy, Time and Performance. In each of those ANOVAs, the only significant effect was Task Card, which was significant at p = 0.044, 0.012 and 0.017, respectively. As shown in **Table 2**, the difficult task card had worse performance on all variables than did the easy task card.

	Accuracy, %	Time, s	Performance, %/s
Easy Task Card	74	754	0.104
Difficult Task Card	58	1073	0.058

Table 2. Results of Pilot Study 1 for Simplified English

Results of Pilot Test 2: Native Chinese Engineering Students

From December 2003 to February 2004, we conducted a pilot test of our methodology before actually collecting data in foreign MROs in China. 40 native Chinese engineering students were recruited from the graduate student pool at the University at Buffalo. We assumed that a Chinese graduate student majoring in Engineering in the United States possessed more knowledge and had a higher ability to use the English language in general than would be typical of maintenance personnel in China. In order to decrease the gap between these two groups, we specified that student participants should have arrived in the United States less than one year ago to be eligible for this experiment. For this pilot test, we used 40 participants in a three factor design (5 Interventions x 2 Simplified English/Not x 2 Task Cards).

For our pilot test group, there were three possible individual variables that might have affected performance: reading level score, years of learning English, and years as an Engineering major. These could have been useful covariates in the analysis of main factors by reducing the expected variability between individual participants. An inter-correlation matrix of these revealed that "Years of Learning English" was significantly correlated with the time to complete the task card comprehension questionnaire (R = 0.498, p = 0.001), and "Reading Level Score" was related to accuracy (R = 0.34, p = 0.032). We decided to consider two covariates: "Year of Learning English" and "Reading Level Score." Note that, as with Pilot Test 1, there was a negative correlation between Accuracy and Time, but here it was not statistically significant (p = 0.091).

We used GLM 3-factor ANOVAs on each performance variable with and without the above covariates, and found statistical significance for Time and Accuracy/Time in both cases. For Time, there was significant effect of Intervention (F(4,20) = 7.77, p = 0.001), and for Accuracy/Time there was significant effect of Task Card (F(1,20)=5.68, p=0.027). As shown in **Table 3**, the easy task card had a worse performance on all variables than did the difficult task card. The results were quite counter-intuitive, with the difficult task card having better performance than the easy one. We suspect that this may have been caused by the potential variability when two versions of each task card were translated into Mandarin. The effects of Simplified English may also have been different for the Mandarin and original versions. In fact, if the "Translation" intervention is eliminated, no terms in the ANOVA are significant.

	Accuracy, %	**Time, s**	**Performance, %/s**
Easy Task Card	66	1364	0.051
Difficult Task Card	72	1202	0.063

Table 3. Results of Pilot Study 2 for Simplified English

All the interventions resulted in decreased accuracy, but shorter time for completion. We did expect that these Chinese graduate students would achieve higher accuracy when comprehending a task card written in their native language. One possible explanation for this is that the aviation maintenance domain is a very specialized domain, so task cards in both English and Chinese were unfamiliar and difficult for the participants, and the advantages of the native language were somehow minimized. If we considered Performance (i.e., Accuracy/Time), all four interventions (except Glossary) resulted in better overall scores than the Control condition, and the Chinese Translation was significantly better than the Control condition at 0.068 vs. 0.500 (T = -7.81, p = 0.004). As a check on the even distribution of participants across Interventions, a one-way ANOVA of Reading Level between Interventions was conducted. As shown in **Table 4**, there were significant differences in Reading Level (F(4,35) = 3.91, p< 0.01), showing that our random assignment did not in fact produce equivalent groups.

	English Reading Level	**Accuracy, %**	**Time, s**	**Performance, %/s**
0 Control	10.3	75	1560	0.050
1. Glossary	11.9	73	1519	0.050
3. Tutoring	8.9	69	1264	0.056
4. Glossary & Tutoring	8.8	61	1027	0.057
2. Chinese Translation	10.6	66	1046	0.068

Table 4. Results of Pilot Study 2 for Interventions

Because "English Reading Level" was significantly different across Interventions, we reconsidered it as a covariate, and ran GLM 3-factor ANOVAs on each performance variable. For performance variables Time and Accuracy/Time, there was not much difference between with and without the new covariate. For Accuracy, with the covariate, the interaction between Intervention and Document English became marginally significant at (F (4, 19) = 2.83, p = 0.054).

Observations

According to our observations, most of the student participants did not utilize the interventions of glossaries, tutoring, or the combination of the above two as much as we had expected. After the experiment, the native Chinese experimenter asked them why they did not utilize the resources. The participants agreed that: "although we do not understand some words, even a sentence here and there, we are still able to answer the comprehension questionnaire; clarifying the meaning of all the details may not necessarily improve our performance, but it will take much longer to finish the task." In fact, this makes sense, as all international students who apply for graduate school in the United States need to submit their scores on the Test of English as Foreign Language (TOEFL), and the Graduate Record Examination (GRE). For non-native English speakers, in order to achieve better scores on the TOEFL and GRE-Verbal tests in a limited time, one key factor is the ability to figure out unknown words, phrases, and even sentences in context. This is a common consensus by non-native English speaking students who have gone through this process. As a result of Pilot Test 2, we have eliminated the combined Glossary and Tutoring condition from our subsequent Asian data collection.

CONCLUSIONS

The main comprehension task took less than half an hour to complete, while the other measures, such as the Reading test and the rating scales, together took another 15 minutes or so. Because many people could be tested together, we were efficient in data collection at the site, and cannot develop accurate timetables for our on-site work in China.

This experiment used a baseline condition of English documents, and then added translation (including the test form), a glossary, a bilingual coach, and a combination of these last two conditions. We used two levels of task card difficulty, each with and without Simplified English. This made a three-factor factorial experiment (Intervention x Difficulty x Simplified English), with the Reading Level score as a covariate. On the samples tested so far, the US and the UK participants obviously had the baseline intervention only, whereas the Chinese-speaking engineering students had all of the interventions. At this stage, any significant effects should be treated with caution, despite the interesting findings on Simplified English and Interventions. These pilot studies are being used for testing the methodology, training the experimenters, and providing an English-speaking baseline condition.

We are planning on doing more data collection using our contacts in China and Taiwan. Data collection at these sites will take place in spring 2004, with other sites in summer 2004.

ACKNOWLEDGEMENT

This work was financed by the Federal Aviation Administration, grant No. 2002-G-025; Contract monitor, William K. Krebs.

REFERENCES

Carver, R. P. (1987). *Technical Manual for the Accuracy Level Test*. REVTAC Publications, Inc.

Chervak, S., and Drury, C. G. (2003). Effects of Job Instruction on Maintenance Task Performance. *Occupational Ergonomics*. Vol.3, Issue 2, 121-131

Chervak, S., Drury, C. G., and Ouellette, J. L. (1996). Simplified English for Aircraft Workcards. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, 303-307.

Dennis, W. (2002). MRO to Grow in China. *Aviation Week and Space Technology*.

Drury, C. G. and Ma, J. (2003). Do Language Barriers Result in Aviation Maintenance Errors? *Human Factors and Ergonomics Society 47th Annual Meeting Proceedings*, Denver, Colorado, October 13-17, 2003.

Drury, C. G. and Ma, J. (2003). *Language Errors in Aviation Maintenance: Year 1 Interim Report*, Reports to William J. Hughes Technical Center, the Federal Aviation Administration under research grant #2002-G-025.

Drury, C.G., Wenner, C., and Kritkausky, K. (2000). Information Design Issues in Repair Stations. *Proceedings of Tenth International Symposium on Aviation Psychology*. Columbus, OH.

MRO in Asia-Pacific Region (2002). Aviation Week's ShowNews Online: www.aviationnow.com. Online resources: *http://www.awgnet.com/shownews/02asia1/mro09.htm* & *Overhaul & Maintenance*.

Patel, S., Drury, C. G.and Lofgren, J. (1994). Design of Workcards for Aircraft Inspection. *Applied Ergonomics*, **25**(**5**), 283-293.

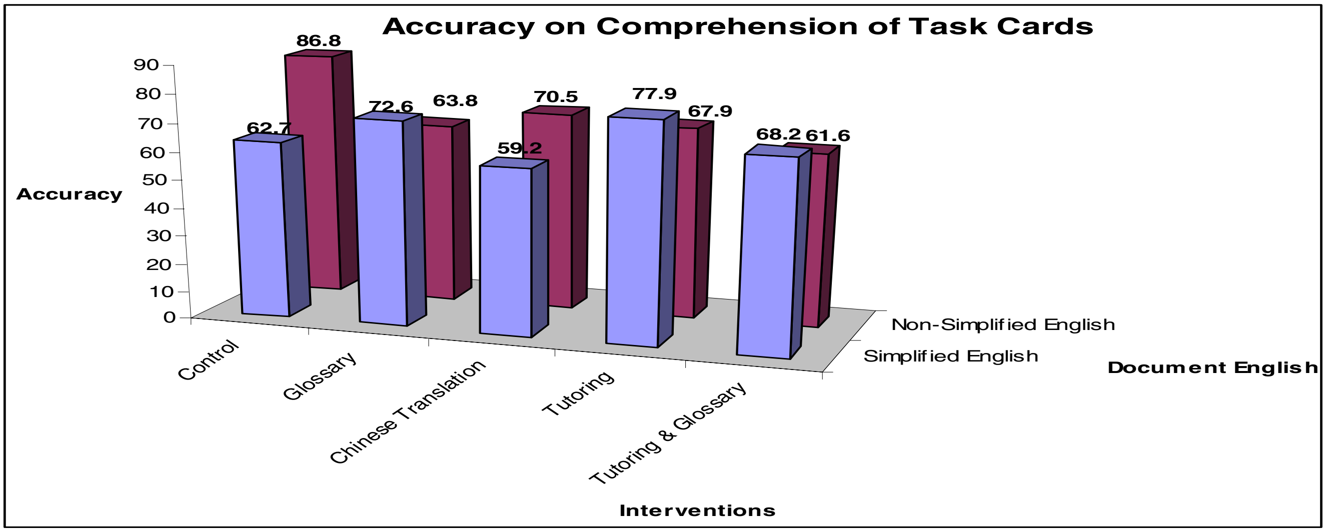

Figure 1. Accuracy on Comprehension of Task Cards